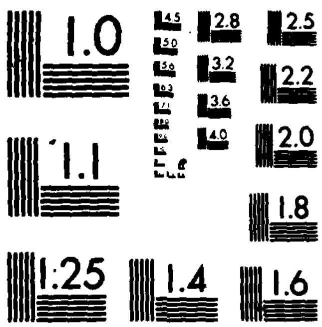
AD-A113 060 CLEMSON UNIV SC DEPT OF MATHEMATICAL SCIENCES F/G 12/1
A FINITE URN MODEL FOR SELECTING THE POPULATION WITH THE LARGES--ETC(U)
JAN 82 K ALAM, M H RIZVI N00014-75-C-0451
UNCLASSIFIED TR-378 NL

END
DATE
FILMED
4-82
DTIC



MICROCOPY RESOLUTION TEST CHART

(12)

A FINITE URN MODEL FOR SELECTING THE POPULATION WITH THE LARGEST α-QUANTILE

*Khursheed Alam and M. Haseeb Rizvi
Clemson University and Sisorex Inc.

DEPARTMENT OF MATHEMATICAL SCIENCES

CLEMSON UNIVERSITY

Technical Report #378

N 136

January, 1982

*This work was supported in part by the U. S. Office of Naval Research under Contract N00014-75-C-0451.

A Finite Urn Model For Selecting The Population With The Largest α-Quantile

Khursheed Alam and M. Haseeb Rizvi
Clemson University and Sisorex Inc.

ABSTRACT

alpha (for =)

Several procedures for ranking populations according to the quantile of a given order have been discussed in the literature. These procedures deal with continuous distributions. This paper deals with the problem of selecting a population with the largest α-quantile from $k \geq 2$ finite populations, where the size of each population is known. A selection rule is given based on the sample quantiles, where the samples are drawn without replacement. A formula for the minimum probability of a correct selection for the given rule, for a certain configuration of the population α-quantiles, is given in terms of the sample numbers.

Key Words: Ranking and Selection, Hypergeometric Distribution, Quantiles, Finite Population.

This work was supported by the U. S. Office of Naval Research under Contract N00014-75-C-0451.

1. Introduction. This paper deals with the problem of selecting a population from amongst several populations whose distribution functions are unknown. In many situations there are certain parameters of the distributions which are of interest, such as, the mean and variance. Then the populations are ranked according to the values of those parameters and the selection is based on their estimated values. It may be required to select, for example, the population with the smallest variance or the population with the largest mean. Whereas, the mean, variance, a scale parameter and a location parameter are parameters of general interest, it is sometimes justifiable to rank the populations according to the quantile of a given order, especially when the standard parameters do not exist. Rizvi, Sobel and Woodworth (1968) have discussed a comparison of populations in terms of α-quantiles. Sobel (1967) and Rizvi and Sobel (1967) have considered the problem of selecting the population with the largest α-quantile and the problem of selecting a subset of $k \geq 2$ populations which includes the population with the largest α-quantile.

The papers mentioned above pertain to large populations. In this paper we consider the problem of selecting the population with the largest α-quantile from $k \geq 2$ finite populations. The populations are sampled by the method of sampling without replacement. A practical situation in which the problem may arise is illustrated by the following example: Suppose that the Department of Education in a certain state is interested in selecting one of several schools in an area to implement

a special training program for exceptionally bright students. As the program involves only exceptionally bright students, the school with the largest 75th quantile score on a special merit examination (SME) may be selected for the training program. A random sample of n students is taken from each school and the selected students are given the SME. The selection of the school for the special training program would depend on the SME scores.

The selection problem is formulated as follows: Let $\pi_1,\ldots,{}_k$ denote $k \geq 2$ finite populations and let N_i denote the size of π_i, $i = 1,\ldots,k$. The numbers $N_1,\ldots,N_k$ are assumed to be known. The members of each population are ranked according to some characteristic value. Let X_{i,m_i} denote the m_ith smallest value of the elements of π_i for $m_i = \alpha N_i$, where α is a given positive fraction. Then X_{i,m_i} represents the α-quantile of π_i. It is assumed that αN_i is integer valued for each $i = 1,\ldots,k$. We shall call the population associated with the largest value of X_{i,m_i} as the best population. It is required to select the best population.

Suppose that a sample of $n_i < N_i$ elements is drawn from π_i without replacement. Let S_i denote the sample α-quantile, that is, the αn_i smallest value in the sample. It is assumed for simplicity that αn_i is integer values for each $i = 1,\ldots,k$. We select the population associated with the largest value of S_i for the best population. We shall call this procedure S.

Let ε be a positive fraction, such that $\varepsilon < \alpha < 1-\varepsilon$. We assume for simplicity that εN_i is integer valued for each $i=1,\ldots,k$. Let C_ε denote a configuration of population quantiles given by

$$(1.1) \qquad X_{j,(\alpha+\varepsilon)N_j} < X_{i,(\alpha-\varepsilon)N_i}, \; j = 1,\ldots,i-1, \; i+1,\ldots,k$$

when π_i is the best population, $i=1,\ldots,k$. This is called a preference zone. It is required that the probability of a correct selection (PCS) for the procedure S should satisfy the relation PCS $\geq$ P* in the preference zone, where P* is a pre-assigned number, such that $\frac{1}{k} < P^* < 1$.

The value of the PCS for the procedure S depends on the sample numbers $n_1,\ldots,n_k$. In the following section we derive a formula for the minimum value of the PCS inside C_ε. The sample numbers needed to satisfy the probability requirement for S are determined from the given formula.

2, Procedure S.

Suppose that π_i is the best population. It is easy to see that the probability of a correct selection for the procedure S, given the configuration C_ε, is minimized when

$$(2.1) \qquad X_{i,(\alpha-\varepsilon)N_i-1} \leq X_{j,1} \leq X_{j,(\alpha+\varepsilon)N_j} < X_{i,(\alpha-\varepsilon)N_i} \leq X_{i,N_i} \leq X_{j,(\alpha+\varepsilon)N_j+1}$$

for $j = 1,\ldots,i-1,i+1,\ldots,k$. Therefore, given C_ε, the minimum probability of a correct selection is equal to

$$(2.2) \qquad \min_{i=1,\ldots,k} P\{S_j < S_i, \; j=1,\ldots,i-1, \; i+1,\ldots,k \mid \pi_i \text{ is best population}\}$$

$$= \min (R_1,\ldots,R_k)$$

where

$$(2.3)\qquad R_i = \sum_{r=0}^{\alpha n_i - 1} \binom{(\alpha-\varepsilon)N_i - 1}{r}\binom{(1-\alpha+\varepsilon)N_i + 1}{n_i - r} \Big/ \binom{N_i}{n_i} \; x$$

$$\prod_{j=1,\ldots,i-1,i+1,\ldots,k} \sum_{r=\alpha n_j}^{n_j} \binom{(\alpha+\varepsilon)N_j}{r}\binom{(1-\alpha-\varepsilon)N_j}{n_j - r} \Big/ \binom{N_j}{n_j}$$

$$= Q(\alpha n_i - 1);n_i,(\alpha-\varepsilon)N_i - 1, N_i) \; x$$

$$\prod_{j=1,\ldots,i-1,i+1,\ldots,k} (1 - Q(\alpha n_j - 1; n_j, (\alpha+\varepsilon)N_j . N_j))$$

and

$$(2.4)\qquad Q(x;n,M,M) = \sum_{r=0}^{x} \binom{M}{r}\binom{N-M}{n-r} \Big/ \binom{N}{n}$$

denotes the cumulative distribution function of the hypergeometric distribution.

The value of the PCS given by (2.2) and (2.3) can be computed from the tables of the hypergeometric distribution prepared by Lieberman and Owen (1961). If N_i is large and n_i is small compared to N_i for each i, then R_i is approximately given by

$$(2.5)\qquad R_i \simeq I_{1-\alpha+\varepsilon}((1-\alpha)n_i + 1, \alpha n_i) \prod_{j=1, j\neq i}^{k} I_{\alpha+\varepsilon}(\alpha n_j, (1-\alpha)n_j + 1).$$

where

$$I_p(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\ (b)} \int_0^p x^{a-1}(1-x)^{b-1}\, dx$$

denotes the imcomplete beta function. On the other hand, if n_i and N_i are large for each i, such that $n_i/N_i \rightarrow \xi_i$, say, where $\xi_i < 2(1 - \max(\frac{\varepsilon}{\alpha}, \frac{\varepsilon}{1-\alpha}))$, then R_i is approximately given by

(see Wise (1954))

$$(2.6)\qquad R_i \simeq I_{w_i}\left((1-\alpha)n_i+1,\ \alpha n_i\right) \prod_{j=1,j\neq i}^{k} I_{w_j}\left(\alpha n_j,(1-\alpha)n_j+1\right)$$

where

$$w_i = 1-\alpha+\frac{2\varepsilon}{2-\xi_i}$$

$$w_j = \alpha+\frac{2\varepsilon}{2-\xi_j}\ .$$

Note that (2.6) reduces to (2.5) for $\xi_1 = \ldots = \xi_k = 0$.

If the k populations are of equal size N, say, we take $n_1 = n_2 = \ldots = n_k = n$, say. Then from (2.2) and (2.3) the minimum value of the PCS is equal to

$$(2.7)\qquad \left(\sum_{r=n}^{\alpha n-1}\binom{(\alpha-\varepsilon)N-1}{r}\binom{(1-\alpha+\varepsilon)N+1}{n-r}\right)\times$$

$$\left(\sum_{r=\alpha n}^{n}\binom{(\alpha+\varepsilon)N}{r}\binom{(1-\alpha-\varepsilon)N}{n-r}\right)^{k-1}\Big/\left(\binom{N}{n}\right)^{k}.$$

Table I below gives the minimum value of n for which PCS $\geq$ P* for α=.50, ε=.05, .10, P*=.75, .95, .99, k=1(1)5 and N = 30, 40, 50, 100, 200, 400. The minimum value of the PCS for the given n is also shown in the table. It is seen from the table that in some cases there is considerable difference between the minimum value of the PCS and the prescribed value P*, due to the discrete nature of the hypergeometric distribution. However, the descrepancy is reduced for large N.

In practice, the given populations would vary in size. Therefore, we consider the value of R_i, given by (2.6). It is easily shown that for large m the beta integral $I_p(\alpha m, (1-\alpha)m+1)$ is increasing (decreasing) in m for $p < (>) \alpha$. Therefore, R_i is an increasing function of $n_1,\ldots,n_k$ when the sample numbers are large. Thus a smallest sample number can prescribed for each population to meet the probability requirement for the procedure S in the general case when the population size varies. In this case it would be interesting to find an optimal distribution of the sample numbers, given $\sum_{k=1}^{k} n_i = n$, say. This is an exercise in linear programming, where we maximize a function $f(\xi_1,\ldots,\xi_k)$ subject to the constraints $\sum_{i=1}^{k} \xi_i N_i = n$ and $0 < \xi_i < 1$, $i = 1,\ldots,k$. Here $f(\xi_1,\ldots,\xi_k) = \min(R_1,\ldots,R_k)$, where R_i is given by (2.6) with the substitution $\xi_\ell N_\ell$ for n_ℓ, $\ell = 1,\ldots,k$.

Remark 1. We observe that the value of the PCS given by (2.2) where R_i is given by (2.5) represents the minimum probability of a correct selection when the samples are drawn with replacement.

Remark 2. It is seen from (2.1) that the value of the PCS is equal to 1 if

$$\frac{n_i}{N_i} > \max\left(1 - \frac{\varepsilon}{\alpha}, \frac{\alpha+\varepsilon}{1-\alpha}\right) \quad , \qquad i = 1,\ldots,k.$$

Therefore, a correct selection is obtained with probability 1 with $n_i < N_i$ for each i.

Acknowledgement. The subject of this article was suggested to the authors by Professor J. Sedransk. Professor M. Sobel helped in the formulation of the problem.

REFERENCES

1. Gibbons, J. D., Olkin, I. and Sobel, M. (1977). Selecting and Ordering Populations - A New Statistical Methodology. Wiley Series in Probability and Statistics.

2. Johnson, N. I. and Kotz, S. (1969). Discrete Distributions in Statistics. Houghton Mifflin Series in Statistics.

3. Lieberman, G. L. and Owen, D. B. (1960). Tables of the Hypergeometric Probability Distribution. Tech. Report No. 50, Stanford University Press, California.

4. Rizvi, M. H. and Sobel, M. (1967). Nonparametric Procedures for Selecting a Subset Containing the Population with the Largest α-quantile. Ann. Math. Statist. (38) 1788-1803.

5. Rizvi, M. H., Sobel, M. and Woodworth, G. G. (1968). Nonparametric Ranking Procedures for Comparison with a Control. Ann. Math. Statist. (39) 2075-2093.

6. Sobel, M. (1967). Nonparametric Procedures for Selecting the t Populations with the Largest α-quantile. Ann. Math. Statist. (38) 1804-1816.

7. Wise, M. E. (1954). A Quickly Convergent Expansion for Cumulative Hypergeometric Probabilities, Direct and Inverse. Biometrika (41) 317-329.

Table I - Minimum Value of n for PCS $\geq$ P* (α = .50)

ε = .05

K		2			3			4			5		
N	P*	.75	.95	.99	.75	.95	.99	.75	.95	.99	.75	.95	.99
30	n	22	28	28	24	28	28	24	28	28	24	28	28
	PCS	.750	1.000	1.000	.809	1.000	1.000	.792	1.000	1.000	.776	1.000	1.000
40	n	28	34	36	30	36	36	32	36	36	32	36	36
	PCS	.788	.955	1.000	.783	1.000	1.000	.821	1.000	1.000	.784	1.000	1.000
50	n	34	44	46	36	44	46	38	44	46	38	44	46
	PCS	.778	.976	1.000	.781	.976	1.000	.808	.976	1.000	.785	.976	1.000
100	n	52	76	84	60	78	86	64	80	86	68	82	86
	PCS	.764	.954	.992	.769	.952	.995	.764	.956	.993	.780	.966	.992
200	n	72	128	150	88	134	154	98	138	156	104	140	156
	PCS	.752	.955	.991	.756	.954	.992	.757	.954	.992	.752	.951	.990
400	n	92	190	244	116	206	254	134	214	258	.46	222	264
	PCS	.755	.950	.990	.753	.952	.991	.756	.950	.990	.754	.952	.991

ε = .10

N		2: .75	2: .95	2: .99	3: .75	3: .95	3: .99	4: .75	4: .95	4: .99	5: .75	5: .95	5: .99
30	n	14	22	24	16	22	24	18	22	24	18	22	24
	PCS	.784	.980	1.000	.794	.972	1.000	.829	.965	1.000	.797	.958	1.000
40	n	16	26	30	18	28	30	20	28	32	22	28	32
	PCS	.771	.958	.994	.760	.972	.992	.769	.965	1.000	.795	.957	1.000
50	n	18	30	36	20	32	36	24	34	38	24	34	38
	PCS	.773	.951	.993	.752	.958	.990	.796	.971	.997	.759	.965	.996
100	n	26	46	60	28	50	62	32	52	62	34	54	64
	PCS	.753	.952	.992	.766	.957	.993	.770	.956	.991	.759	.959	.993
200	n	26	62	88	34	68	92	38	72	94	42	76	96
	PCS	.755	.951	.991	.755	.953	.991	.753	.953	.991	.751	.955	.990
400	n	28	76	144	38	84	120	44	90	126	50	94	130
	PCS	.751	.953	.990	.765	.953	.990	.750	.953	.991	.765	.951	.991

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: N 136	2. GOVT ACCESSION NO.: AD-A113060
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): A Finite Urn Model for Selecting the population with the largest α-quantile	5. TYPE OF REPORT & PERIOD COVERED
	6. PERFORMING ORG. REPORT NUMBER: TR #378
7. AUTHOR(s): Khursheed Alam and M. Haseeb Rizvi	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Clemson University, Dept. of Mathematical Sciences, Clemson, South Carolina 29631	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 365-049
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Code 434, Arlington, Va. 22217	12. REPORT DATE: January, 1982
	13. NUMBER OF PAGES: 11
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Ranking and Selection, Hypergeometric Distribution, Quantiles, Finite Population.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Several procedures for ranking populations according to the quantile of a given order have been discussed in the literature. These procedures deal with continuous distributions. This paper deals with the problem of selecting a population with the largest α-quantile from $k \geq 2$ finite populations, where the size of each population is known. A selection rule is given based on the sample quantiles, where the samples are drawn without replacement. A formula for the minimum probability of a correct (over)

selection for the given rule, for a certain configuration of the population α-quantiles, is given in terms of the sample numbers.